DOCUMENT RESUME

ED 232 691 IR 050 339

AUTHOR Bookstein, Abraham
TITLE Recent Developments in the Theory of Information Retrieval. Stockholm Papers in Library and Information Science.
INSTITUTION Royal Inst. of Tech., Stockholm (Sweden). Library.
REPORT NO TRITA-LIB-6019
PUB DATE Dec 82
NOTE 29p.; Based on a seminar presented at the Royal Institute of Technology Library, November 10, 1982, funded by DFI (the Swedish Delegation for Scientific and Technological Information).
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Check Lists; *Information Retrieval; *Online Systems; *Probability; Relevance (Information Retrieval); Search Strategies
IDENTIFIERS Boolean Logic; *Fuzzy Set Theory; *Weighted Term Searching

ABSTRACT

Recently considerable attention has been given in the online information retrieval literature to techniques for producing a weighted output of documents in response to a request. One approach tries to maintain the form of and relationships among requests as they appear in current Boolean logic-based systems, while extending it to permit a weighted output. It is based on the mathematics of fuzzy-set theory, which assigns each potential member of a set a degree of membership between zero and one with intermediate values denoting partial membership in the set. Another approach is based on the mathematics of probability. It represents requests by sets of terms, and, by means of feedback information, assigns a weight to each term. Documents are ordered by the sum of weights of the terms in the request that match those in the documents. This paper provides an overview of both approaches and their advantages and disadvantages. It also outlines a basic theory of information retrieval, discusses the characteristics and problems of using Boolean logic in information retrieval, and lists areas for continuing research. An appendix describing mathematical constraints on fuzzy-complements and a 21-item bibliography conclude the paper. (Author/ESR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

REPORT TRITA-LIB-6019 DECEMBER 1982

Stockholm Papers in Library and Information Science

ED232691

RECENT DEVELOPMENTS IN THE THEORY OF INFORMATION RETRIEVAL*

Abraham Bookstein
Graduate Library School
University of Chicago

* Based on a seminar presented at the Royal Institute of Technology Library, 10 November 1982, during a stay as visiting professor, funded by DFI (The Swedish Delegation for Scientific and Technological Information)

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. Schwarz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR050339

Stockholm papers in library and information science
TRITA-LIB-6019, RITL, Stockholm 1982
ISSN 0346-9042

SUMMARY

Recently considerable attention has been given in the information retrieval literature to techniques for producing a weighted output of documents in response to a request. Although weighting mechanisms have existed for some time, the more recent approaches have attempted to develop a theoretically sound basis to guide the implementation. Two rather different approaches have been taken. The first tries to maintain the form of and relationships among requests as they appear in current Booleanlogic based systems, while extending it to permit a weighted output; it is based on the mathematics of Fuzzy-set theory, which will be briefly described. The others is based upon the mathematics of probability. It represents requests by sets of terms and, by means of feedback information, assigns a weight to each term. Documents are ordered by the sum of weights of the terms in the request that match those in the documents. The paper provides an overview of both approaches, and notes some of the interesting problems that are currently receiving attention.

TABLE OF CONTENTS

SUMMARY 3

I. INTRODUCTION 5
II. BOOLEAN RETRIEVAL 6
III. FUZZY-SET RETRIEVAL 11
IV. PROBABALISTIC RETRIEVAL 16
V. OUTLINE OF THE THEORY 21

APPENDIX 24
Constraints on fuzzy-complements

REFERENCES 26

I. INTRODUCTION

The rate of development in the technology of information retrieval over the last decade can only be described as startling. The increased capacity and reduced costs of computers and storage devices, improved communications capability, and the proliferation of very large machine readable data bases have made the notion of on-line retrieval, only recently futuristic, commonplace. The recent appearance and spread of micro-computers, as well as the promise of new storage technologies, encourage the belief that, at least for the near future, the explosive growth in automated retrieval will continue virtually unchecked.

If we turn to the retrieval strategies being used by current information retrieval systems (IRS), however, we find remarkably little variation from those used twenty years ago. Yet the last few years has seen theoretical development in information retrieval as rapid and fundamental as that occurring in its technology. There is a great deal of enthusiasm within the research community about the direction this work is taking: the believe is widespread that we are developing coherent and satisfying ways of understanding the information retrieval process, and that the procedures we are creating show real promise for improving the performance of IRSs. The purpose of this paper is to introduce the two theoretical developments that have received the most attention recently, and to direct the reader to pertinent papers.

Perhaps the best way to begin is by contrasting the new approaches with those most heavily used by current systems. In the next section I will make explicit the assumptions that underlie most of the current systems, since the newer approaches begin by rejecting one or more of these assumptions. An explicit statement of these assumptions should also serve the purpose of encouraging us to stand back and look more critically at the principles on which our procedures depend. In subsequent sections, the theory of fuzzy-set retrieval and probabalistic retrieval will be described.

II. BOOLEAN RETRIEVAL

The predominant approach taken by current operating IRSs is based on Boolean logic. In such systems, each document is represented by an index record of descriptors, and requests are statements of the conditions an index record must satisfy for the document it describes to be retrieved. These requests are constructed by connecting index terms with the logical connectives AND, OR and NOT. A precise statement of Boolean retrieval is given in (1).

Most complex systems are constructed with some ideal in mind. This is true of Boolean IRSs as well. To understand this ideal, it is helpful to ask ourselves, what should be required for a Boolean system to operate perfectly. The following conditions are most important:

Assumption 1: Completeness of Indexing Vocabulary.

For any collection of documents, all aspects of that collection that will in the future be useful for distinguishing which of those documents are relevant for a patron from those not relevant can be represented with a finite vocabulary of index terms.

Assumption 2: Binary Index Term Assignment.

To describe a document for retrieval purposes, it is adequate simply either to assign an index term or not. It is not necessary for retrieval to indicate a degree of association of a term with a document. The index record of a document, then, is simply a list of all the index terms assigned to that record.

Assumption 3: Indexing Coherency.

For a Boolean system to operate without error, it must be possible for the system to assign index terms to documents unambiguously: that is, all

indexers must agree with each other regarding the assignment of index terms, and, more importantly, these assignments must agree with the terms future users of the system would have assigned. Ultimately, indexing a document serves only one purpose: to provide a mechansim for future retrievals of that document. A future user of the system formulates his request with a conceptualization of how he would have indexed the documents he would want to see and those he would not want to see; he implicitly assumes that the indexers originally assigning terms to documents have done so in the same way he would have, for otherwise relevant documents will be missed or nonrelevant documents retrieved. Thus, in a sense, an indexer assigns terms as a surrogate for future users of the system, and trying to anticipate their needs and requests. In an ideal system, the match will be perfect.

Assumption 4: Dichotomous Relevance Scale.

Given any information need, it is adequate to conceive the collection as being made up of only two types of documents: those that would satisfy that need and those that would not.

Assumption 5: Congruence of Boolean Statements with Information Needs.

Because of assumption 3, the user has a good idea of the characteristics that would cause a document to be indexed by a particular set of index terms. Assumption 5 states that it is possible for the user to make use of this knowledge by translating his information need into a request in Boolean format, based on these index terms, in such a way that those documents that would satisfy his information need and only those documents would satisfy the conditions of the request. For this to be true, not only must the index vocabulary be adequate in scope and precision to allow the user to distinguish documents (Assumption 1), but the Boolean apparatus must be adequate, in fact as well as in principle, for representing his information need in terms of this indexing vocabulary: the configuration of index terms distinguishing relevant from non-relevant documents must be describable in Boolean format.

Two divisions of the document collection are possible: the first, which is a system task, separates documents satisfying the request from those that do not; the second is a user responsibility - it separates relevant from non-relevant documents. Assumption 5 says that requests can always be constructed in such a way that these decompositions are identical.

Several consequences follow from these assumptions. Two deserve comment:

Consequence 1: Document and logical term independence.
These are technical considerations mentioned because parallel statements are commonly made in the approaches soon to be discussed. Document independence refers to the assumption that all documents satisfying the conditions of the request are in fact relevant and should be retrieved; the appropriateness of retrieving a document depends only on the properties of that document itself as reflected in its index record, and not on the presence of other documents in the collection. Logical term independence is more subtle. What I have in mind here is that Boolean requests tend to be formulated with terms the user considers pertinant to his need, and only these terms influence the retrieval decision. Typically, the retrieved set of document will contain other terms as well. It is a consequence of our assumptions that those terms in a document's index record but not explicitly in the request have no influence on the appropriateness of retrieving that document.

Consequence 2: Set theoretic formulation.
We have described the Boolean retrieval mechanism in terms of conditions satisfied by index records. But a well known parallel interpretation exists, formulated in set theoretic terms. This description will be particularly useful when we describe the fuzzy-set approach.

To describe Boolean retrieval in set theoretic terms, we first note that because of Assumption 2, the indexing process implicitly defines a set of documents associated with each indexing term: viz. the set of documents indexed by that term. It is easy to see, then, that a document will satisfy the

Boolean condition (A AND B), for index terms A and B, if it is in the set that is the intersection of the sets associated with A and B. Similarly, the Boolean statement (A OR B) is associated with the union of the two sets, and NOT (A) with the complement of the set associated with A. Since arbitrary Boolean expressions are constructed out of simple expressions of this type, we can conclude that given any Boolean request, the retrieval process can be defined in terms of set operations carried out on the sets of documents that are associated with the index terms comprising the request.

Recently, these assumptions, and the systems based on them, have been severely criticised (see e.g. 2). Of course, we don't expect that these assumptions would exactly hold in practice, any more than any complex system could be an exact expression of the ideal inspiring it. The point is that it has been suggested that in this case, the ideal so far diverges from reality that restricting ourselves to Boolean systems has become a very serious impediment to providing adequate service, especially considering the alternatives that are available to us today. In particular, the following specific complaints may be made:

1. Many people have difficulty understanding and using Boolean logic and find the consequences of Boolean logic counterintuitive. This is a direct attack on assumption 5: it claims that Boolean logic is not, in fact, the best mechanism for expressing an information need. Consider, for example, a request of the form A AND B AND C AND D (say: Information Retrieval AND Fuzzy-set Theory AND Probabalistic Modelling AND Overview). A document (this one, for example) indexed by all of these terms would be retrieved in response to this request. However, if three of these terms were present but the fourth absent, the document would not be retrieved - the document would be treated by the system in exactly the same way it would treat a document in which no terms were present. This result follows, of course, from the two value relevance scale discussed above: the system is incapable of making any distinctions finer than relevant or not-relevant, and since the absence of a single term means the document fails to satisfy the formal conditions of the

request, the document is considered to be not relevant. The system is properly following the rules of Boolean logic, but is acting counter to our intuition. Similarly, consider a request of the form A OR B OR C (for example: Magnetic Disk storage OR Bubble memory OR Video Disk Storage Devices). Here, whether a document is indexed by a single one of these terms or all of them, it is retrieved with equal strength. However, many people feel that a document indexed by all of these terms is in fact likely to satisfy the request better and that this should be recognized by the IRS.

2. Related to the first criticism is the complaint that ordinary Boolean retrieval systems do not provide a ranked output - again an attack on Boolean retrieval's two value relevance scale. A ranked output is believed helpful to the user as well as to the system itself.

3. It is widely known that indexers do not agree with each other about how a document should be indexed (3,4,5), a violation of assumption 3. Some current thinking has it that this is a characteristic intrinsic to the information retrieval process: it should be explicit in our models of IRS and not viewed as an imperfection that may be eliminated by a finer tuning of the system.

4. More generally, some researchers express a sense of unease that much of current practice has evolved under the pressure of daily operation and without developing a conceptualization regarding what information retrieval (IR) is and what it should be doing. They ask, how should we be thinking about the IR process? These people argue that such a comprehension is essential if we are to develop new systems that improve on the performance of those now in existence.

Over the last few years, attempts have been made to conceptualize what an IRS is and, specifically, to do so in such a way that responds to the criticisms I noted regarding conventional Boolean retrieval. Much of this work is formal and makes heavy use of mathematics; nontheless, many of the ideas are not difficult to understand - indeed, simply noting them often

leaves the impression of their being obvious. But when acted upon they do have far reaching consequences and provide a mechanism for designing detailed systems.

Of the two approaches I am about to describe, the first is based on the theory of fuzzy-sets. This approach essentially operates within a Boolean framework, and seeks to generalize it by introducing a ranking mechanism, rather than by replacing it altogether. It accepts assumptions 1, 3 and 5 and modifies assumptions 2 and 4.

The probability approaches appear completely to reject the Boolean approach and to introduce a completely new retrieval procedure. These approaches reject assumptions 3 and 5 of Boolean retrieval, but they do implicitly accept as an assumption the two value relevance scale of Boolean retrieval. In these approaches, the object is the take advantage of clues present in a document's index record to assess the probability that the document is relevant; documents are then ranked according to these probabilities.

III. FUZZY-SET RETRIEVAL

The defining property of a conventional set is that it is possible to decide unambiguously whether any object in the universe of discourse is in the set or not in the set. However, there are many set-like like entities whose membership is not so clearly determined: an example favoured by fuzzy-set theorists is the "set" of beautiful women, though the "set" of documents relevant to a request would be an example more appropriate to this discussion. In these cases, some items may be clearly within the set, and others clearly not in the set. However, for many items membership is not clear. The notion of a "fuzzy-set" was first proposed by Lofti Zadeh (6) in 1965 to generalize the concept of a traditional set in such a way that encompasses these situations. To define a fuzzy-set, each potential member of the set is assigned a degree of membership between zero and one, with zero denoting non-membership and one, full membership. Intermediate values denote partial membership in the set.

Recently, a number of papers have appeared in the IR literature that have sought to use the fuzzy-set idea as a means of retaining the essential features of Boolean retrieval while making it more flexible (7, 8, 9, 10, 11, 12). See, however, (13) for a criticism of this approach . These papers accept assumption 1 as before. Assumption 2 is changed to permit indexers to assign a value between zero and one (see Appendix for a more general statement) to any index term in a document to indicate partial applicability of the term. Thus once every document has been indexed in this way, we have in fact associated a fuzzy-set with each index term: the degree of membership of a document in an index term's set is simply the degree with which that index term was initially assigned to the document.

Similarly, assumption 4 is modified to allow users to represent their degrees of satisfaction with a document on a scale between zero and one. This implicitly forms the fuzzy-set of "relevant documents". Assumption 3 is accepted, though indexing coherency now includes agreement on the assignment of weights. Assumption 5 is also accepted: requests (in most treatments) are simple Boolean formulations; however, it is now a fuzzy-set that is retrieved. That is, just as in ordinary Boolean retrieval we retrieve the set of documents that satisfy the conditions of the request, so we now also retrieve such a set, except that the set is a fuzzy-set, in which each document is assigned a weight indicating its degree of membership. Thus we recognize that some documents satisfy the request more than others and, making the translation justified by an extension of Boolean assumption 5, these documents are presumed to be desireable for the requestor to the same degree that they satisfy the request: that is, the set of retrieved documents is identical to the set of relevant documents, where both sets are now fuzzy-sets.

Two characteristics of fuzzy-set retrieval, as it has usually been presented, should be noted, as they contrast with the approach we shall discuss next.

1. The fuzzy-set approach is intrinsically deterministic. The weights that are assigned to documents represent degrees of relevance, and these are conceived of as being made with no (or a negligeable) degree of uncertainty.

2. The fuzzy-set approach concentrates on the logic of retrieval, just as does the traditional Boolean approach. As is well known, Boolean algebra was developed to formalize the laws of "logical thinking," and requests formulated for a Boolean IRS are assumed to observe these laws; indeed, it is this assumption that permits us to represent Boolean retrieval in set theoretic terms. A fundamental characteristic of fuzzy-set theory is that it also observes almost all of the axioms of Boolean algebra, and thus closely mimics the properties of the conventional systems.

With the above background, we can now describe in more detail how a fuzzy IRS would function. As noted above, requests take the exact same form as they do in ordinary Boolean systems, thought attempts have been made to permit weighting of the terms making up the request (8,9). The Fuzzy-set approach differs most strikingly from the traditional Boolean approach in that when a document is indexed, terms are not simply assigned or not assigned, but each assigned term is also given a weight indicating the system's assessment of the extent to which that term describes the document. Once all the documents are indexed in this way, we have in fact created a fuzzy-set for each index term. To retrieve documents, we perform set operations parallel to those we would use in an ordinary Boolean based system, except that the set operations now take a form determined by fuzzy-set theory. This can be described as follows:

1. Intersection: The weight of a document in the intersection of sets A and B is the minimum of its weight in A and its weight in B. For example, if a document is in set A with strength .8 and in B with a strenght .3, it would be in the intersection of the two sets with a weight of .3. (If the only weights permitted by the system are zero and one, then a document will be in the intersection of two sets with

a weight of zero unless it has a weight of one in both sets - in this way, ordinary set intersection is a special case of fuzzy-set intersection, wherein the system permits only zero or one membership values.)

2. Union: The weight with which a document is in the union of two sets, A and B, is the maximum of its weights in A and its weight in B. Continuing with the above example, the document would be in the union of A and B with a weight of .8. Union also generalises the notion for ordinary sets.

3. Complementation: If a document is in a set, A, with weight a then it is in the complement of A with weight 1 - a. For zero/one weights, this reduces to the ordinary set complement.

Although these rules appear rather arbitrary, those for union and intersection are in fact forced on us if we wish to preserve the rules of logic referred to above. The rule for complementation is not so strictly imposed, but, as shown in the appendix, no freedom is in fact possible if we restrict the possible degrees of membership to lie in a finite set. Thus, after performing the prescribed set operations, the result is a fuzzy-set of documents - that is, a set of documents, with a weight assigned to each. These weights are then used to rank the documents presented to the user: they are the system's assessment of the extent to which each document is relevant to the user.

We are now in a position to assess the strengths and weaknesses of the fuzzy-set approach:

Advantages: 1 This approach is very close to that of the traditional Boolean retrieval with which we are already familiar. The arguments that would support the Boolean approach would most likely apply equally well here. The fuzzy-set approach is ideal for someone committed to the logic of Boolean retrieval, but is dissatisfied by its associated dichotomous relevance scale. It is thus a conservative extension of current systems.

2 It follows that requests can be manipulated and simplified just as they could in a conventional Boolean system.

3 Documents could be ranked: The system is not restricted to a simple yes or no decision with regard to each document. Similarly, indexers are given greater flexibility: given a term whose applicability an indexer is not sure of, he has an alternative to simply selecting it or rejecting it.

Disadvantages: 1 If the critics of Boolean retrieval are correct, that Boolean logic is not an adequate way of representing information needs, then fuzzy-set systems, which are firmly embedded within the Boolean framework, offer no real improvement. The fuzzy-set approach shares the assumptions of conventional Boolean retrieval, simply generalizing them to present ranked output.

2 The capacity for ranking is not sensetive to all the terms in the request. For example, if a document is in set A to degree .8 and in set B to degree .1, then it is in the intersection of the two sets to degree .1. But this value would not change even if the first value were reduced from .8 to .2: intersection (AND) reflects only the minimum value. A similar comment applies to the union of two sets. As noted before, this is a necessary consequence of our being committed to the laws of traditional logic; the probability approach, by rejecting these rules, permits a more sensitive weighting scheme.

3 Some critics reject altogether the value of fuzzy-sets as a generalization of traditional sets. For some of these, it may be a response to the novelty of the fuzzy-set concept and to the manner in which some advocates of fuzzy-sets have presented their case.

IV. PROBABALISTIC RETRIEVAL

The probabalistic retrieval approach rejects the Boolean framework altogether. As before, documents are represented by records giving features of the documents. Most commonly, index terms are used, but any property of a document is a candidate for inclusion, for example citations, name of journal, age, etc. Quantitative features, taking arbitrary numerical values, can also be useful. An example is the number of times a term or concept appears in the document (15). The important thing about each feature is that it potentially will provide a clue, even if imperfect, as to the relevance of a document to a request - by this I mean that for some of the requests we are likely to receive, the feature has a different distribution pattern in documents relevant to the request than it does in documents that are not relevant. In the following, we will, for simplicity, assume binary features: those which either appear in a document or not. If we know such a feature is present in a particular document, this knowledge gives us some indication as to whether the document should be retrieved.

Generally we are looking at a large number of clues. If so, it is quite possible that each clue in itself gives rather weak evidence regarding the relevance of the document to a request. However, collectively, they can make a strong case for retrieving or not retrieving a document. Thus whereas in Boolean retrieval we assumed that features were unambiguiously associated with documents and placed our emphasis on expressing the logical relation of features in relevant documents, now the presence of features is probabalistically related to the requests and we are seeking methods of combining and balancing evidence indicating the relevance of a document for a patron. The idea that the uncertainty in indexing should be explicit in our models of IR was first proposed by Maron and Kuhns (14). Modern probabalistic IR exploits these ideas.

In a probabilistic IRS, then, each document is represented by a record containing potential dues as to its relevance for future requests. These clues may be assigned by a human indexer; more in keeping with the flavor of this approach, they will be assigned automatically - for example, index terms

may be all significant works (or word stems) occurring in the abstract of a document. A thesaurus may be introduced for vocabulary control. A user of the system will, in turn, provide the system with a request which takes the form of list of features he thinks would characterize the sort of documents he would like to see. These requests, like index records, take the form of simple lists - no Boolean connectives are used. The second major break with Boolean retrieval involves the retrieval mechanism. First we note that as with the earlier Boolean systems, probabalistic systems conceptualize documents as simply being either relevant to a request or not relevant. No intermediate degrees of relevance are recognized - largely because of the technical difficulties involved (see, however, (15)). However, given the available clues, the system is able to compute the probability that a document is relevant, and the objective of the system is to provide each user with a list of documents ordered according to the likelihood that they are relevant. The ideas that probabilities could be used to design a detailed retrieval mechanism; that the retrieval mechanism can, in some sense, provide optimal retrieval; and the initial development of the pertinant mathematics appeared independently in papers written by Bookstein (16), Robertson (17) and Robertson and Sparck-Jones (18); the last of three papers (18) was the first to suggest the use of feedback information. More recently Bookstein extended the theory to incorporate feedback information as an intrinsic component of the model (19) and generalized the model to include non-binary relevance assessments (15).

I can now describe how such a system would look in operation. The retrieval process is made up of the following stages:

1. Upon receiving a request the system assigns a weight to each feature in the request. This initial weight could simply be one for each feature, or it could be a value assigned by the patron; more commonly the frequency of occurrence of the feature in the total corpus is the basis for this assignment. For example, - log (freq) has provided good

results - features occurring frequently are deemed less likely to be useful for discriminating documents. Croft and Harper (20) were the first to provide a heuristic justification for the success of the log(freq) weights; his argument itself depends on the probabilistic reasoning on which the rest of the approach is based and provides an important unity for this retrieval approach.

2. For each document, we compare the terms or features by which the document was indexed with those in the request. We add the weights of the matching features; this is the weight assigned to that document.

3. The patron is then presented with a list of documents ordered by these weights. He evaluates each document and informs the system on the documents which are relevant.

4. On the basis of this feedback, the weights are reevaluated in a manner prescribed by the theory. This could involve removing some terms that originally appeared in the request, or introducing into the request features the user didn't think of, but which seem to discriminate between relevant and non-relevant documents. In an SDI system, this ability to modify profiles could form the basis for an evolutionary system, in which the system automatically responds to changes in technology and in patron interests. It also permits - indeed, may require - including negative weights for some terms that appear to be particularly associated with nonrelevant documents. This possibility is often overlooked by systems employing ad hoc weighting schemes.

5. We then use the new weight to retrieve another set of documents. The procedure is then repeated until the patron is satisfied that no more useful documents are being generated. In the extreme case, the patron could interact with the system through a terminal; the system would output a small number of documents and immediately take advantage of the feedback information to retrieve the next group of documents; in the limiting case, a user could be presented with documents one at a time.

Before indicating some of the technical assumptions, it will be useful to indicate some of the strengths and weaknesses of this approach.

Advantages: 1 The approach has a firm theoretical base. I should point out here that weighting as such is not new. What is original here is the attempt to provide a theoretical justification for the approach, and to base the algorithms by which weights are assigned and modified on theoretical models. The body of theory used relies very heavely on the mathematics of probability. For each feature it tries to evaluate the probability of that feature appearing in a relevant document as compared to that of the same feature appearing in a non-relevant document. If these probabilities are the same, for example, that feature conveys no information that can distinguish a relevant from a non-relevant document and no weight is assigned to that feature. If the probabilities are very different, then the appearance of that feature constitutes strong evidence regarding the relevance or non-relevance of those documents in which it appears and its appearance should be taken seriously: it should be given a large weight. Ultimately, what is computed is the probability that a document is relevant, given the evidence that is available (the features by which a document is indexed) and it is this probability on which the documents are ranked. There is now a considerable body of theory suggesting how this should be done.

A comment on the interpretation of these weights is appropriate here, since these weights take on a very different meaning from those used in fuzzy-set retrieval. In fuzzy-set logic, no probabilities are involved and the weights represent the actual degree of relevance of each document to the request. Now we are no longer distinguishing different levels of relevance: documents are

considered in these probability models to be relevant or not. The weight represents the probability that the document is relevant. The weighting mechanism then tells us how much the presence of each feature contributes to this probability, and how to take feedback information into account in changing these weights. These two interpretations are reconciled in (15).

2 This approach isn't burdened with Boolean connectives; in this sense it is simpler than the Boolean and fuzzy-set approaches.

3 Ranking of documents is a natural consequence of the methods employed, and the weights by which the documents are ranked are sensitive to every feature.

4 Methods have been developed for taking advantage of feedback information. This may also be possible in the fuzzy-set approach, but has not yet been done - the existence of the logical connectives makes this difficult, although index records may, perhaps, be modified with experience from feedback.

Disadvantages: 1 As generally presented, all logical connections between terms are lost. These probably are of value, however, even if the full Boolean apparatus is rejected as inappropriate.

2 Increased storage capability and computation are required.

3 The detailed weights are model dependent. We have no operational experience to indicate how valid these models are and whether the costs of implementation can be justified by the improvement (if any) in performance. What would be particularly valuable, now, are experiments carried out on live systems, with actual requests and realistic evaluations.

V. OUTLINE OF THE THEORY

Any IRS that presents documents to a patron one at a time, or simulates this by giving him an ordered set of documents, can be conceptualized as a sequential process. After presenting n documents to a patron, and perhaps receiving the patron's evaluations of these documents, the task before the system is to determine which document to present next (or, perhaps, to terminate the search).

The methodology most heavily used is that of decision making under uncertainty. This approach recognizes that information retrieval is, first of all, a decision making process - each retrieval is a decision to bring a particular document to a patron's attention rather than any other. As in other decision making situations, there are costs associated with making decisions. These costs may be measured in units of money, time, frustration, etc. In practice, these costs may be very difficult to estimate; however, often we can derive useful results simply by being able to assume the meaningfulness of these costs, without actually having to assign specific values. Our intention, then, is to make that decision that incurs the smallest cost.

If our decisions are made under conditions of uncertainty, however, we do not know ahead of time how a patron will respond to a particular document. Thus we need some mechanism for dealing with uncertainty. We assume here that

a) These uncertainties can in principle be quantified
b) These quantities can be manipulated using the rules of probability

The actual probabilities used are derived from evidence available in index records and, possibly, feedback information; theoretical models are used to connect the two. Ultimately, we compute the expected cost of retrieving each document, and retrieve that document next whose expected cost is the smallest. This criterion can be shown to be equivalent, for the simpler cost models, and with a binary relevance scale, to retrieving next that document whose probability of relevance is highest. This criterion, after some manipulation, can be reduced to the following criterion function:

$$f_R(x_1, x_2, \ldots, x_n; h) / f_{R'}(x_1, x_2, \ldots, x_n; h);$$

here:

a) $x_1, x_2, \ldots, x_n$ are features describing documents; the x_i's take values for each document: for example, one if a specific index term is in the index record, zero otherwise.

b) h represents the feedback history at the point the decision is being made.

c) f_R and $f_{R'}$ are probability distributions. They give the probability of the features taking specific values in relevant (R) and non-relevant (R') documents, respectively.

At this point, simplifying assumptions have to be made. The most common is that of term independence, although recently efforts have been made to introduce more realistic models (21). The independence assumption asserts that:

$$f_R(x_1, \ldots x_n; h) = f_R(x_1; h)\, f_R(x_2; h) \ldots f_R(x_n; h),$$

with a parallel equation holding for $f_{R'}$. Using this simplification, it is then possible to transform the criterion function to the following form:

$$\sum \log \left[(f_R(x_i; h)/f_R(0_i; h)) / (f_{R'}(x_i; h)/f_{R'}(0_i; h)) \right]$$

where $f_R(0_i; h)$ as the probability that the i-th feature take its null value (e.g. the index term not appearing), and the sum is over all non-null features in the document being evaluated. It should be noted that the simple summation of weights is a consequence of the independence assumption. At this point, the weighting procedures diverge depending on the functions chosen to model the probabilities.

Areas of continuing research: The following problems and tasks remain:

1. These models should be tested under realistic conditions.

2. Alternatives to the independence assumption that are effective and computationally efficient should be developed.

3. Much of the experimental research uses binary valued features. We should investigate the usefullness of multiple valued features, for example, the number of occurrences of a term in a document.

4. We should explore the value of features not appearing in a request itself; ref (19) suggests how this can be done.

5. Techniques need to be developed to estimate the parameters arising in our models.

6. Extensions of the theory to include a non-binary relevance scale should be attempted.

APPENDIX

Constraints on fuzzy-complements

In the text, as in most discussions of fuzzy-sets, we assumed our membership function took values from the continuum running between zero and one. This is at once too general and too restrictive for information retrieval: - too general in that in an operating system, almost certainly document relevance (or index term assignments) will be chosen from a finite set of values. For example, we might extend the binary Boolean case by introducing a single additional value, "undecided"; or, values can be taken from the set (0, .1, .2, ..., .9, 1); - too restrictive in that values outside the range zero to one may be more convenient. Indeed, verbal rather than numerical designations may be most appropriate.

In the following development, then, I will simply assume that membership values are taken from a finite set, the "membership set", whose members are simply ordered: given two different members, we can always say one or the other definitely is the "larger". We first note that this assumption has no impact on the rules of fuzzy-intersection and union: the derivations in (10) assume only such a ordering exists and do not depend on their taking values in the interval (0,1). Below, we shall examine the implications of this assumption for fuzzy-complement. To do this, we will state two axioms of ordinary set theory that we wish to preserve:

1. For a set A, $(A^C)^C = A$. (In terms of logic, NOT (NOT p) is logically equivalent to p, for p a statement);

2. For sets A and B, $(A \cup B)^C = A^C \cap B^C$ (de Morgan's law).

Suppose, then, that a point has a membership value x in a set A. We denote by f (x) its value in the complement of A. Our task is to explore how much freedom we have in choosing f. We shall show that for the case of a finite membership set, no freedom at all exists. We first note that since a point with membership value x in A has membership value f(x) in A^C, it will have

membership value f (f (x)' in $(A^C)^C$. But $(A^C)^C$ = A; thus f must obey f (f(x)) = x: f is its own inverse. Also, only one value can be taken by f onto any value a. For suppose b and c both map onto a. Then, since f(f(x)) = x, we have both f(a) = f (f (b)) = b and f (a) = f (f(c)) = c.

Thus b and c must be identical. Another important consequence of our assumptions is that if a < b, then f(b) < f(a). This follows from de Morgan's laws and the rules for fuzzy intersection and union. For suppose, indeed, a< b. Then, if a point p has membership value a in set A and b in set B, its membership value in $A \cup B$ is b, and its membership value in $(A \cup B)^C$ is f(b). But $(A \cup B)^C = A^C \cap B^C$. Thus, f (b) = Min (f (a), f(b)), that is, a< b implies f(b)< f(a) (we showed above that equality was not a possibility).

With these preliminary results we can now state our basic result: under the conditions stated, f is unique. The proof is by induction. We shall prove the assertion for membership sets with an odd number of values. A similar proof serves for even sized membership sets. Suppose, then, that the membership set is of size 2n-1. For n = 1, the set has only one point, a, so f (a) = a and f trivially is uniquely determined.

Suppose that we have shown the assertion to be true for all values of n<N. Consider the case for N: suppose a is the lowest value in the membership set and c the highest. I assert that f (a) = c (and thus f (c) = a). For let f(a) = b, where b ≤ c. Then f (c) ≤ f (b). But f (b) = a and a is the smallest member of the set. Thus f (c) = f (b), and thus c = b, as was to be shown. Further, by the uniqueness property of f, no other member of the set could be assigned to a or c. Thus, evaluating f on the membership set becomes equivalent to evaluating it for the set after the values a and c are removed - i.e, for a set of size (2N-1) - 2 = 2(N-1)-1. But the induction hypothesis has it that for n = N-1, f is uniquely determined. Thus it is uniquely determined for N as well.

Combining this case with the case for even sized sets, we see that we are correct in concluding that f is unique: as is the case with union and intersection, fuzzy complement is determined by the axioms of set theory.

REFERENCES

1. Bookstein, A and Cooper, W.S: "A General Mathematical Model for Information Retrieval System", Library Quarterly, 46 (1976) 153-67.

2. Cooper, W.S.: "Maximum Entropy Principle and its Application to Probabalistic Information Retrieval Design". Interim Report. School of Library and Information Studies, University of California, Berkeley, California, June, 1981.

3. Zunde, P. & Dexter, M.E.: "Indexing Consistency and Quality", American Documentation, 20 (1969) 259-267.

4. Hurwitz, E.K: "A Study of Indexer Consistency", American Documentation, 20 (1969) 92-4.

5. Cooper, W.S.: "Is Interindexer Consistency a Hobgoblin?", American Documentation, 20 (1969) 268.

6. Zadeh, L.A.: "Fuzzy Sets", Information and Control, 8 (1965) 338-53.

7. Tahani, V.: "A Fuzzy Model of Documentation Retrieval Systems". Information Processing and Management, 12 (1976) 177-88.

8. Bookstein, A.: "Fuzzy Requests: An Approach to Weighted Boolean Searches". Journal of the American Society for Information Science, 31 (1980) 240-47.

9. Buell, D.A. & Kraft, D.H.: "A Model for a Weighted Retrieval System". Journal for the American Society for Information Sciences, 32 (1981) 211-16.

10. Bookstein, A.: "A Comparison of two Systems of Weighted Boolean Retrieval". Journal of the American Society for Information Science, 32 (1981) 275-79.

11. Bookstein, A.: "Weighted Boolean Retrieval". Proceedings of the Joint BCS/ACM Conferences on Information Retrieval, Cambridge, England, 1980.

12. Radecki, T.: "Mathematical Model of Information Retrieval System Based on a Concept of Fuzzy Thesaurus". Information Processing and Management, 12 (1976) 313-18.

13. Robertson, W.E.: "On the Nature of Fuzz: A Diabribe". Journal of the American Society for Information Science, 29 (1978) 711-6.

14. Maron, M.E & Kuhns, J.L: "On Relevance, Probabalistic Indexing and Information Retrieval". Journal of the Association for Computing Machinery, 7 (1960) 216-244.

15. Bookstein, A.: "Outline of a General Probabalistic Retrieval Model". To appear.

16. Bookstein, A. & Swanson, D.: "A Decision Theoretic Foundation for Indexing". Journal of the American Society for Information Science, 26 (1975) 45-50.

17. Robertson, S.E.: "The Probability Ranking Principle in IR". Journal of Documentation, 34 (1977) 294-304.

18. Robertson, S.E. & Spark-Jones, K.: "Relevance Weighting at Search Terms", Journal of the American Society for Information Science, 27 (1976) 129-146.

19. Bookstein, A.: "Information Retrieval: A Sequential Learning Process". Journal of the American Society for Information Science, 34 (1983).

20. Croft, W.B. & Harper, D.J.: "Using Probabalistic Models of Document Retrieval Without Relevance Information". Journal of Documentation, 35 (1979) 285-295.

21. Van Rijsbergen, C.J.: "A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval". Journal of Documentation, 33 (1977) 106-119.